A genome-wide investigation of microsatellites mismatches and the association with body mass among bird species

Haiying Fan [Corresp., 1] , **Weibin Guo** [1]

[1] Department of Ecology, College of Life Sciences, Wuhan University, Wuhan, China

Corresponding Author: Haiying Fan
Email address: fanhaiying1989@126.com

Mutation rate is usually found to covary with many life history traits of animals such as body mass, which has been readily explained by higher number of mutation opportunities per unit time. Although the precise reason for the pattern is not yet clear, to determine the universality of this pattern, we test whether life history traits impact another form of genetic mutation, the motif mismatches in microsatellites. Employing published genome sequences from 65 avian species, we explore the motif mismatches patterns of microsatellites in birds on a genomic level and assess the relationship between motif mismatches and body mass in a phylogenetic context. We have found that small-bodied species have a higher average mismatches and we suggest that higher heterozygosity in imperfect microsatellites lead to the increase of motif mismatches. Our results obtained from this study, suggest that a negative body mass trend in mutation rate may be a general pattern of avian molecular evolution.

1 **A genome-wide investigation of microsatellites mismatches and the association with body**

2 **mass among bird species**

3 **Haiying Fan[1] & WeibinGuo[1]**

4 [1]Department of Ecology, College of Life Sciences, Wuhan University, Wuhan 430072, China

5 Corresponding author: Haiying Fan, email: fanhaiying1989@126.com

6 Department of Ecology, College of Life Sciences, Wuhan University, Wuhan 430072, China

24

25

26

27

28

29

30

31

32

33 **Abstract**

34 Mutation rate is usually found to covary with many life history traits of animals such as body

35 mass, which has been readily explained by higher number of mutation opportunities per unit time.

36 Although the precise reason for the pattern is not yet clear, to determine the universality of this

37 pattern, we test whether life history traits impact another form of genetic mutation, the motif

38 mismatches in microsatellites. Employing published genome sequences from 65 avian species,

39 we explore the motif mismatches patterns of microsatellites in birds on a genomic level and

40 assess the relationship between motif mismatches and body mass in a phylogenetic context. We

41 have found that small-bodied species have a higher average mismatches and we suggest that

42 higher heterozygosity in imperfect microsatellites lead to the increase of motif mismatches. Our

43 results obtained from this study, suggest that a negative body mass trend in mutation rate may be

44 a general pattern of avian molecular evolution.

Introduction

45

46   It has long been recognized that the molecular evolutionary rates always covary with many life

47   history traits of animals. Numerous studies have documented a negative relationship between the

48   rate of molecular evolution and body mass (Nabholz, Glémin & Galtier, 2008; Bromham, 2011;

49   Amos & Filipe, 2014), where genes in small-bodied species are likely to evolve faster than those

50   in large-bodied species. This has been readily explained by higher number of mutation

51   opportunities per unit time (generation length hypothesis, Li et al., 1996) or higher mutation

52   probability in a round of DNA replication due to higher metabolic rate (metabolic rate hypothesis,

53   Mindell et al., 1996) in small-bodied species. Although the precise reason for the pattern is not

54   clear at present, to determine the universality of this pattern, we need to study additional form of

55   genetic mutation besides mitochondrial DNA or nuclear 'genes' which are most frequently used.

56   The first to consider is the fastest evolving components of the genome such as microsatellites.

57

58   Microsatellites, also known as simple sequence repeats (SSRs), are tandem repeats of simple

59   nucleotide motifs, which have wide coverage in eukaryotic and prokaryotic genomes (Tóth,

60   Gáspári & Jurka, 2000; Ellegren 2004; Adams et al., 2016). One feature of microsatellites is that

61   they have a high mutation rate (10^{-7} to 10^{-3} mutations per locus per generation), leading to high

62   heterozygosity and extensive length polymorphisms (Kruglyak et al., 2000). It has long been

63   assumed that the major cause of variation of microsatellite repeats is replication slippage

64   (Kornberg et al., 1964; Bhargava & Fuentes, 2010), which will increase or decrease repeat copy

65   numbers in microsatellites. Specifically, when it creates a loop in one of strands, a slippage error

66   occurs. If the loop is formed in the replicating strand, it will introduce an insertion. If the loop is

67   in the template strand, a deletion will emerge. Several mathematical models of microsatellite

68   evolution have been proposed to represent the mutation processes of microsatellites, such as

69   stepwise mutation model (SMM) of Ohta & Kimura (1973), which suggests that mutation in

70   microsatellite loci occurs by one repeat unit at a time.

71

72   Many studies on microsatellites have explored the frequencies, abundance and polymorphism of

73   microsatellites in the genomes (Wang et al., 2014; Qi et al., 2015; Adams et al., 2016). Few, if

74   any, have correlated these microsatellite characters to the life history traits of a species.

75   Specifically, microsatellites are hypothesized to experience a life cycle: start short (birth) and

76   expand predictably due to mutation bias (expansion) until they become unstable and either

77   collapse or degrade through internal point mutations (contraction and death) (Chambers &

78   MacAvoy, 2000; Buschiazzo & Gemmell, 2006). Life history traits of species are expected to

79   have an influence on the life cycle —'birth', expansion, and 'death'—of microsatellites in the

80   genome (Amos & Filipe, 2014). For example, in smaller species, higher mutation rate allows the

81   'birth' and expansion of microsatellites faster, due to higher mutation rate and slippage rate.

82   Since the death rate is lower than the birth rate, microsatellites tend to accumulate in the genome

83   (Buschiazzo & Gemmell, 2006). In that, the smaller species harbour a higher frequency of

84   microsatellites across the genome, which has been proved in mammals (Amos & Filipe, 2014).

85

86   It is well known that except for repeat copy number variation, a microsatellite (e.g.

87   ATATATATAT) also suffers from nucleotide substitutions and insertion/deletion mutations,

88   hence becoming imperfect (e.g. ATATATCATAT: AT repeat with an insertion of C). Perfect

89   and imperfect microsatellites are thus defined. It has been found that genomes possess a

90   relatively small but significant number of imperfect microsatellites (Brinkmann et al., 1998).

91   Mismatch variation of imperfect microsatellites is critical for their maintenance in the genome

92   and imperfect microsatellites are more stable compared to perfect microsatellites since the

93   former is less prone to slippage mutation (Sturzeneker et al., 1998). Several previous studies

94   have already revealed the genome-wide motif imperfection pattern among species (e.g. Behura &

95   Severson, 2015). Nevertheless, our understanding of motif mismatches in imperfect

96   microsatellites is still very limited and their correlation with life history traits remains to be

97   revealed and explained.

98

99   In this study, we used 65 avian genome sequences, employing SciRoKo (Kofler, Schlötterer &

100 Lelley, 2007) to search SSRs in the whole genome. We chose avian genomes for this study

101 because microsatellites have been widely used in population genetics of bird species, yet the

102 pattern of microsatellites mismatches in birds is still not well understood, mostly owing to the

103 lack of avian genomic information. With the advance of whole genome sequencing, evolution of

104 microsatellites is attracting attention from researchers. With the genome-wide microsatellites

105 data in hand, we presented the first detailed comparative study of microsatellites, aiming to

106   reveal the patterns of motif mismatches across different bird species and to help understand the

107   relationship with life history traits.

108

109   **Materials & Methods**

110   *Genome sequences and body mass*

111   We downloaded fasta files of the 65 avian genomes from NCBI

112   and GigaDB (http://dx.doi.org/10.5524/101000). These avian species represent nearly all of the

113   major clades of living birds. We compiled data from the original and secondary references and

114   the world-wide web about the mean body mass of adult males and females (Table S1). If a mean

115   value was not provided for a species, we took the median of the range. Where separate body

116   masses were given for males and females, the average value of the masses was calculated.

117

118   *Identification of microsatellites*

119   We searched microsatellites in each genome sequences using SciRoKo 3.4, a simple sequence

120   repeats (SSRs) identification program (Kofler, Schlötterer & Lelley, 2007), with the default

121   parameters (minimum score = 15 and mismatch penalty = 5) in the mismatched modes. In

122   addition, we used different parameters to search SSRs (minimum score = 15 and mismatch

123   penalty = 3, minimum score = 10 and mismatch penalty = 5) considering changing in parameters

124   would affect the results of this study. Specially, the motif mismatches refer to the number of base

125   mismatches of an imperfect microsatellite compared with its idealized perfect counterpart. For

126   example, the string TACTACTAGTACTAC, is a trinucleotide repeat with five repeats and, by

127   comparison with its idealized perfect counterpart (consensus repeat), it has a mismatch of 1. The

128   number of mismatches of each microsatellites as well as their length for each genome was used

129   for different comparative analyses across the species.

130

131   *Statistical analysis*

132   In this study, we used phylogenetic generalized least-squares regression (PGLS) (Freckleton,

133   Harvey & Pagel, 2002) implemented in the package 'ape' (Paradis, Claude & Strimmer, 2004) to

134   control shared ancestry (for the script used, see Figure S1). We used the evolutionary tree of the

135   48 bird species estimated by Jarvis et al. (2014) as a backbone topology, and used the

136   phylogenetic information provided by Jetz et al. (2012) to add the remaining 17 species (for the

137   resulting phylogeny, see Figure S2). In order to achieve the statistical requirements for linearity

138   and normality, adult average mass were log10-transformed prior to analysis. Average

139   mismatches was reciprocal transformed. GC content was arcsine square root transformed.

140

141   Firstly, we computed some basic statistics on characteristics of microsatellite loci in 65 bird

142   genomes (Table S2, S3, S4, S5). Secondly, to better understand the occurring of motif

143   mismatches in bird genomes, we determined the frequency of microsatellites of 20 bp that either

144   lacks mismatch or harbours one mismatch for each species (Table S6). 20bp was used because

145   that the average length of perfect microsatellites of 65 birds is 20bp. Then we computed the ratio

146   of imperfect (mismatch = 1) repeats frequency to perfect (mismatch = 0) ones. Then, we

147   employed a PGLS, treating the ratio of imperfect (mismatch = 1) to perfect (mismatch = 0)

148   repeats as a dependent variable and body mass as an explanatory variable. Thirdly, we explored

149   whether or not the extent of motif mismatches is related to genomic abundance of imperfect

150   microsatellites. We first calculated the probability of per-site mismatch (the total number of

151   mismatches divided by total lengths of all loci) in each genomes. Then the expected number of

152   mismatches was determined based on the length and compared with the observed number of

153   mismatches in each imperfect microsatellite (Table S7). The first paired sample t-test was

154   conducted between the numbers of microsatellites harbouring more mismatches than expected

155   and that of carrying fewer mismatches than expected. The second paired sample t-test was

156   performed between imperfect repeats that have a length of at least 30 bp and have either <3 or ≥ 3

157   mismatches (Table S8). Finally, to test whether differences of mismatches in imperfect

158   microsatellites link with body mass, we fitted a PGLS analysis with average mismatches of

159   imperfect SSRs as dependent variable and body mass as a predictor. The average mismatches of

160   imperfect SSRs in individual genomes was estimated as the sum of mismatches divided by the

161   number of imperfect microsatellites (Table S1). Average mismatches was used because it

162   indicates the mismatches in an 'average' imperfect SSR. For controlling the probability that GC

163   content will have a potential influence on microsatellite mismatches, we added it to the models

164   as a predictor variable. Taking di-, tri-, tetra-, penta- and hexamers as the five classes of repeats,

165   we repeated the PGLS analysis in each repeat type. Since the mutations in the mononucleotide

166   repeats tend to cause the emergence of a new motif of other repeat type, we excluded it from our

167   analysis. All statistical analyses were conducted with R 3.1.2 (R Development Core Team 2014).

168

169 **Results**

170

171 *(a)   Characteristic of microsatellite loci in 65 avian species*

172 In total, 11803896 SSR loci with a minimum length of 15 bp were identified from 65 avian

173 genome assemblies, and were classified into mono-, di-, tri-, tetra-, penta- and hexanucleotide

174 SSRs according to the motif length (Table S2). Among these, mononucleotide SSRs are the most

175 abundant (42.3%) type, followed distantly by tetra- (18.8 %) and pentanucleotide SSRs (17.1%)

176 (Table S2; Figure S3). The SSR abundance composition and SSR density of the birds varies

177 greatly among species, with the maximum value in *Anas platyrhynchos* (416040 counts; 376.49

178 counts/Mb) and the minimum value in *Melopsittacus undulatus* (81643 counts; 73.07 counts/Mb).

179 Additionally, the SSR abundance composition are predicted by genome size ($\beta \pm SE = 1.56 \pm$

180 0.64, $t = 2.44$, $P = 0.017$, $R^2 = 0.09$).

181

182 *(b)   Frequency of imperfect microsatellites in bird genomes*

183 The number of imperfect microsatellites varies among the birds species and the imperfect repeats

184 account for 15–27% of all microsatellites searched from the genome assemblies of the 65 bird

185 species as shown in Table S3. The imperfect repeats represented less than 0.2% of the genome

186 sequence in most of these birds except four species (*Anas platyrhynchos, Calypte anna, Columba*

187 *livia, Picoides pubescens*) (Figure S4). The data in Table S3 and Figure S4 shows that the

188 frequency of imperfect microsatellites in bird genomes appears substantial variation among these

189 species. It was observed that *Anas platyrhynchos* has a higher percentage of imperfect

190 microsatellites than other bird species. Moreover, the proportion of imperfect repeats varies

191 differentially among species, to some extent, depending on the motif size of microsatellites.

192 Specifically, the paired sample *t*-test results indicated the di-, tri- and hexanucleotide SSRs have

193 an increased rate of motif mismatches compared with all other types of motif size (Table S4).

194 Furthermore, it seems that this pattern is conserved among different avian species.

195

196 *(c)   The occurring of motif mismatches*

197 Imperfect microsatellites are longer than perfect microsatellites in each species (37 vs 20 bp; $t =$

198 33.334, $df = 64$, $p < 0.001$; Table S5). The PGLS analysis revealed that the small-bodied species

199   has a higher ratio of imperfect (mismatch = 1) to perfect (mismatch = 0) repeats of 20bp than

200   large-bodied species ($\beta \pm SE$ = -0.006 $\pm$ 0.002, t = 2.86, P = 0.006, R^2 = 0.12).

201

202   *(d)   The accumulation of mismatches in imperfect microsatellites and genomic abundance*

203   The paired sample t-test revealed that the microsatellites harboring mismatches higher than

204   expected has significantly lower abundance than that carrying mismatches lower than expected

205   (13381 versus 25138 counts; t = 22.651, df = 64, p < 0.001; Table S7), implying that the

206   imperfect microsatellites who containing more mismatches have lower abundance in the genome.

207   We also found that loci with three or more number of mismatches are less common than that

208   have less than three mismatches (7364 vs 18361 counts; t = 11.316, df = 64, p < 0.001; Table S8).

209

210   *(e)   correlation between body mass and average motif mismatches*

211   We found that on a whole genome scale, the average body mass accounts for 28.2% of the

212   variation in average mismatches of imperfect SSRs (Table 1, Figure 1). Body mass also has a

213   significantly negative correlation with microsatellites mismatches in five motif length classes

214   (Table 1, Figure 2). This negative correlation remains significant when adding GC content to the

215   regression models. Inclusion of GC content only enhances the model's explanatory power

216   slightly except in tetra- and pentanucleotide SSRs. When we used different parameters including

217   minimum score 15 and mismatch penalty 3 and a minimum score of 10 and mismatch penalty 5

218   to search microsatellites in the genomes, the results of repeated analyses were highly consistent

219   (Table S9). This confirmed that our observations were not influenced by the search parameters of

220   microsatellites.

221

222   **Discussion**

223

224   In the present study, we did genome-wide search of microsatellites using SciRoKo with the same

225   parameters to ensure that the program can search all possible microsatellites with the same

226   probability for every genome. Microsatellites search results showed that the frequency of

227   microsatellites varies extensively among species. We have also found a positive relationship

228   between microsatellites abundance and genome size among 65 bird species, which is consistent

229   with earlier studies (e.g. Hancock, 1996). After providing a general description of the basic

230   characteristics of microsatellites, we particularly focused on comparing the motif mismatches of
231   imperfect microsatellites to body mass across bird species in a phylogenetic context.
232
233   We found a negative relationship between body mass and the ratio of frequency of imperfect
234   repeats (mismatch = 1) to perfect (mismatch = 0) ones with the same length 20bp among the
235   species. Moreover, it is known that mutations in microsatellites shorter than a critical length are
236   generally gain or loss of single repeat units which cannot disturb the repeat tract (Buschiazzo &
237   Gemmell, 2006). Whereas when it reached a critical length, mismatch was introduced, a perfect
238   microsatellite became imperfect. Here, our result implied that the introduction of motif
239   mismatches in imperfect microsatellites is significantly associated with the nature of point
240   mutation in microsatellites. In small-bodied species, since more perfect microsatellites suffer
241   from the introduction of mismatches due to the higher mutation rate, larger number of imperfect
242   microsatellites relative perfect ones can be observed.
243
244   We observed that the microsatellites harbouring mismatches higher than expected have lower
245   abundance than that carrying mismatches lower than expected. Consistent with this result, we
246   also found that the microsatellites ≥30 bp and >3 mismatches have lower abundance than that
247   ≥30 bp and <3 mismatches, indicating that mismatches of motifs is a key determinant leading to
248   a paucity of long imperfect microsatellites in the genome. That is to say, mismatches would
249   stabilize the repeat array and impede the further expansion. When the extent of mismatches
250   reached saturation point, the repetition pattern is interrupted, leading the microsatellites to
251   degeneration and death. (Taylor, Durkin & Breden, 1999; Harr & Schlotterer, 2000; Yamada et
252   al., 2002; Vowles & Amos, 2006). Although the exact details of death is still poorly understood,
253   the relative number of older mismatches in an 'average' microsatellite is likely to reflect the
254   mutability during its lifetime. It can be further confirmed by the finding that the average
255   mismatches of imperfect SSRs decreases with increasing body mass.
256
257   Our results that higher average mismatches of imperfect SSRs in small-bodied species support a
258   correlation between mutation rate and life history traits. The pattern is usually explained by a
259   generation length model, where smaller species evolve faster due to higher number of mutation
260   opportunities per unit time (Li et al., 1996). In addition, body mass might affect the mutation rate

261   through a link with metabolic rate and/or body temperature, which can directly change the

262   mutation probability in a round of DNA replication (Mindell et al., 1996). Apart from these two

263   key hypotheses, a rising hypothesis which proposes mutation rates are influenced by

264   heterozygosity (Amos, 2010) can better explain the intrinsic correlation of motif mismatches

265   with body mass. Smaller species have larger number of imperfect microsatellites which has been

266   demonstrated by our data ($\beta \pm SE$ = -0.008 $\pm$ 0.002, t = 4.008, $P < 0.001$; R^2 = 0.203). Meanwhile,

267   more heterozygous sites at these imperfect microsatellites can be expected. Recognition and

268   'repair' of heterozygous sites during synapsis will cause additional rounds of DNA replication

269   which in turn provide more opportunities for mutations (Amos, 2011) and introduce more motif

270   mismatches at imperfect microsatellite sites. Therefore, a negative relationship between body

271   mass and motif mismatches can be observed. We suggest that heterozygote instability hypothesis,

272   which is supported by increasing evidence (Drake, 2007; Masters et al., 2011; Amos, 2013;

273   Amos, 2016), could provide a potential link between body mass and motif mismatches. However,

274   further studies are needed in order to examine carefully whether homologous imperfect

275   microsatellites are generally more prone to introduce mismatches in smaller species with a detail

276   comparison between sister species.

277

278   **Conclusions**

279

280   In conclusion, the present study is the first effort to explore the motif mismatches patterns of

281   microsatellites in birds on a genomic level. Our results obtained from this study provide support

282   for the long-standing correlation between mutation rate and life history traits and suggest that a

283   negative body mass trend in mutation rate may be a general pattern of avian molecular evolution.

284

285   **Acknowledgements**

286   We thank Xin Lu, Hongtao Xiao, Guoyue Zhang, Changcao Wang, Qingchen Zhang and

287   Juanjuan Rao for data collection, statistical advice and insightful discussions. We also thank

288   William Amos and Andrew Clarke for helpful suggestions and two anonymous referees for

289   comments on earlier versions of this manuscript.

290

291   **References**

292 **Adams RH, Blackmon H, Reyes-Velasco J, Schield DR, Card DC, Andrew AL,**

293   **Waynewood N, Castoe TA. 2016.** Microsatellite landscape evolutionary dynamics across

294   450 million years of vertebrate genome evolution. *Genome* **59**:295–310 DOI 10.1139/gen-

295   2015-0124.

296 **Amos W. 2010.** Heterozygosity and mutation rate: evidence for an interaction and its

297   implications. *BioEssays* **32**:82–90 DOI 10.1002/bies.200900108.

298 **Amos W. 2011.** Population-specific links between heterozygosity and the rate of human

299   microsatellite evolution. *Journal of Molecular Evolution* **72**:215–221 DOI 10.1007/s00239-

300   010-9423-2.

301 **Amos W. 2013.** Variation in heterozygosity predicts variation in human substitution rates

302   between populations, individuals and genomic regions. *PLoS ONE* **8**:e63048 DOI

303   10.1371/journal.pone.0063048.

304 **Amos W, Filipe LNS. 2014.** Microsatellite frequencies vary with body mass and body

305   temperature in mammals, suggesting correlated variation in mutation rate. *PeerJ* **2**:e663

306   DOI 10.7717/peerj.663.

307 **Amos W. 2016.** Heterozygosity increases microsatellite mutation rate. *Biology Letters* **12**:

308   20150929 DOI org/10.1098/rsbl.2015.0929.

309 **Behura SK, Severson DW. 2015.** Motif mismatches in microsatellites: insights from genome-

310   wide investigation among 20 insect species. *DNA Research* **22**:29–38 DOI

311   10.1093/dnares/dsu036.

312 **Bhargava A, Fuentes FF. 2010.** Mutational dynamics of microsatellites. *Molecular*

313   *Biotechnology* **44**:250–266 DOI 10.1007/s12033-009-9230-4.

314 **Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B. 1998.** Mutation rate in human

315   microsatellites: influence of the structure and length of the tandem repeat. *American*

316   *Journal of Human Genetics* **62**:1408–1415 DOI 10.1086/301869.

317 **Bromham L. 2011.** The genome as a life-history character: why rate of molecular evolution

318   varies between mammal species. *Philosophical Transactions of The Royal Society B-*

319   *Biological Sciences* **366**:2503–2513 DOI 10.1098/rstb.2011.0014.

320 **Buschiazzo E, Gemmell NJ. 2006.** The rise, fall and renaissance of microsatellites in eukaryotic

321   genomes. *BioEssays* **28**: 1040–1050 DOI 10.1002/bies.20470.

322   **Chambers GK, MacAvoy ES. 2000.** Microsatellites: consensus and controversy. *Comparative*
323   *Biochemistry and Physiology B-Biochemistry & Molecular Biology* **126**: 455–476 DOI
324   10.1016/S0305-0491(00)00233-9.

325   **Drake JW. 2007.** Too many mutants with multiple mutations. *Critical Reviews in Biochemistry*
326   *and Molecular Biology* **42**:247–258 DOI 10.1080/10409230701495631.

327   **Ellegren H. 2004.** Microsatellites: simple sequence with complex evolution. *Genetics* **5**: 435–
328   445 DOI 10.1038/nrg1348.

329   **Freckleton RP, Harvey PH, Pagel M. 2002.** Phylogenetic analysis and comparative data: a test
330   and review of evidence. *American Naturalist* **160**:712–726 DOI 10.1086/343873.

331   **Hancock JM. 1996.** Simple sequences and the expanding genome. *BioEssays* **18**:421–425 DOI
332   10.1002/bies.950180512.

333   **Harr B, Schlotterer C. 2000.** Long microsatellite alleles in Drosophila melanogaster have a
334   downward mutation bias and short persistence times, which cause their genome-wide
335   underrepresentation. *Genetics* **155**: 1213–1220.

336   **Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B,**
337   **Howard JT, Suh A, Weber CC, da Fonseca RR, Li JW, Zhang F, Li H, Zhou L, Narula N,**
338   **Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldon T,**
339   **Capella-Gutierrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B,**
340   **Warren WC, Ray D, Green RE, Bruford MW, Zhan XJ, Dixon A, Li SB, Li N, Huang**
341   **YH, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV,**
342   **Wirthlin M, Schneider MPC, Prosdocimi F, Samaniego JA, Velazquez AMV, Alfaro-**
343   **Nunez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce**
344   **M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong ZJ, Zeng YL, Liu**
345   **SP, Li ZY, Liu BH, Wu K, Xiao J, Yinqi X, Zheng QM, Zhang Y, Yang HM, Wang J,**
346   **Smeds L, Rheindt FE, Braun M, Fjeldsa J, Orlando L, Barker FK, Jonsson KA, Johnson**
347   **W, Koepfli KP, O'Brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR,**
348   **Glenn TC, McCormack J, Burt D, Ellegren H, Alstrom P, Edwards SV, Stamatakis A,**
349   **Mindell DP, Cracraft J, Braun EL, Warnow T, Jun W, Gilbert MTP, Zhang GJ. 2014.**
350   Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**:
351   1320–1331 DOI 10.1126/science.1253451.

352 **Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012**. The global diversity of birds
353   in space and time. *Nature* **491**:444–448. DOI 10.1038/nature11631.

354 **Kofler R, Schlötterer C, Lelley T. 2007.** SciRoKo: a new tool for whole genome microsatellite
355   search and investigation. *Bioinformatics* **23**:1683–1685 DOI
356   10.1093/bioinformatics/btm157.

357 **Kornberg A, Bertsch LL, Jackson JF, Khorana HG. 1964.** Enzymatic synthesis of
358   deoxyribonucleic acid: XVI. Oligonucleotides as templates and the mechanisms of their
359   replication. *Proceedings of the National Academy of Sciences of the United States of*
360   *America* **51**:315–323 DOI 10.1073/pnas.51.2.315.

361 **Kruglyak S., Durrett R., Schug M. D., Aquadro C. F. 2000.** Distribution and abundance of
362   microsatellites in the yeast genome can be explained by a balance between slippage events
363   and point mutations. *Molecular Biology and Evolution* **17**:1210–1219 DOI
364   org/10.1093/oxfordjournals.molbev.a026404.

365 **Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996.** Rates of nucleotide
366   substitution in primates and rodents and the generation-time effect hypothesis. *Molecular*
367   *Phylogenetics and Evolution* **5**:182–187 DOI 10.1006/mpev.1996.0012.

368 **Masters BS, Johnson LS, Johnson BGP, Brubaker JL, Sakaluk SK, Thompson CF. 2011.**
369   Evidence for heterozygote instability in microsatellite loci in house wrens. *Biology Letters*
370   **7**:127–130 DOI 10.1098/rsbl.2010.0643.

371 **Mindell DP, Knight A, Baer C, Huddleston CJ. 1996.** Slow rates of molecular evolution in
372   birds and the metabolic rate and body temperature hypotheses. *Molecular Biology and*
373   *Evolution* **13**:422–426 DOI 10.1093/oxfordjournals.molbev.a025601.

374 **Nabholz B, Glémin S, Galtier N. 2008.** Strong variations of mitochondrial mutation rate across
375   mammals – the longevity hypothesis. *Molecular Biology and Evolution* **25**:120–130 DOI
376   10.1093/molbev/msm248.

377 **Ohta T, Kimura M. 1973.** A model of mutation appropriate to estimate the number of
378   electrophoretically detectable alleles in a finite population. *Genetics Research* **22**: 201–204
379   DOI 10.1017/S0016672308009531.

380 **Paradis E, Claude J, Strimmer K. 2004.** APE: analysis of phylogenetics and evolution in R
381   language. *Bioinformatics* **20**: 289–290 DOI 10.1093/bioinformatics/btg412.

382   **Qi WH, Jiang XM, Du LM, Xiao GS, Hu TZ, Yue BS, Quan QM. 2015.** Genome-wide

383   survey and analysis of microsatellite sequences in bovid species. *PLoS ONE* **10**: e0133667

384   DOI 10.1371/journal.pone.0133667.

385   **Sturzeneker R, Haddad LA, Bevilacqua RAU, Simpson AJG, Pena SDJ. 1998.** Polarity of

386   mutation in tumor-associated microsatellite instability. *Human Genetics* **102**:231–235 DOI

387   10.1007/s004390050684.

388   **Taylor JS, Durkin JMH, Breden F. 1999**. The death of a microsatellite: a phylogenetic

389   perspective on microsatellite interruptions. *Molecular Biology and Evolution* **16**:567–572

390   DOI 10.1093/oxfordjournals.molbev.a026138.

391   **Tóth G, Gáspári Z, Jurka J. 2000.** Microsatellites in different eukaryotic genomes: survey and

392   analysis. *Genome Research* **10**:967–981 DOI 10.1101/gr.10.7.967.

393   **Vowles EJ, Amos W. 2006.** Quantifying ascertainment bias and species-specific length

394   differences in human and chimpanzee microsatellites using genome sequences. *Molecular*

395   *Biology and Evolution* **23**: 598–607 DOI 10.1093/molbev/msj065.

396   **Wang JF, Qi, HG, Li L, Zhang GF. 2014.** Genome-wide survey and analysis of microsatellites

397   in the Pacific oyster genome: abundance, distribution, and potential for marker development.

398   *Chinese Journal of Oceanology and Limnology* **32**:8–21 DOI 10.1007/s00343-014-3064-z.

399   **Yamada NA, Smith GA, Castro A, Roques CN, Boyer JC, Farber RA. 2002.** Relative rates

400   of insertion and deletion mutations in dinucleotide repeats of various lengths in mismatch

401   repair proficient mouse and mismatch repair deficient human cells. *Mutation Research-*

402   *Fundamental and Molecular Mechanisms of Mutagenesis* **499**: 213–225 DOI

403   10.1016/S0027-5107(01)00282-2.

404   **Figure and Table legends**

405

406   **Table 1: Result for the relationship between average mismatches and body mass fitted in**

407   **PGLS analyses.**

408

409   **Figure 1: Regression scatterplot of the inverse of the average mismatches of imperfect**

410   **SSRs on the log of body mass in whole genome scale.**

411

412   **Figure 2: Regression scatterplot of the inverse of the average mismatches of imperfect**

413   **SSRs in five classes of repeat type on the log of body mass.**

414

415   **Figure S1: The script for performing the PGLS analyses in R.**

416

417   **Figure S2: Avian phylogeny used in this study.** The phylogeny is presented in Nexus format

418   and can be drawn using any standard tree-drawing package such as TreeView.

419

420   **Figure S3: SSR characteristics for motif sizes 1-6 with minimum 3 repeats in 65 birds.**

421

422   **Figure S4: Imperfect microsatellites as the percentage of genome size of 65 birds.** The

423   abbreviated bird names are shown in the x-axis and the percentages are shown in the y-axis.

424

425   **Table S1: A list of the 65 avian species and average adult body mass, GC content, average**

426   **mismatches of imperfect microsatellites on the whole genome.** Species names are abbreviated

427   with four letters; first letter represents the genus name and last three letters represent the species

428   name. For *Pelecanus crispus* and *Podiceps cristatus*, we use Pecri and Pocri separately. Mono-,

429   di-, tri-, tetra-, penta- and hexa- are microsatellite types.

430

431   **Table S2: Microsatellite abundance in 65 birds.**

432

433   **Table S3: Number of imperfect microsatellites in different avian genomes and the**

434   **percentage of imperfect microsatellites in the corresponding genome.**

435

436   **Table S4: Motif length and percentage of imperfection of microsatellites in different species.**

437   **The paired sample *t*-test results are shown below the data tables.**

438

439   **Table S5: Average length of imperfect microsatellites compared to perfect microsatellites.**

440

441   **Table S6: Genomic abundance of microsatellites having a length of 20 bp that either lack**

442   **mismatch (perfect motifs) or have exactly one mismatch in each locus across species.**

443

444   **Table S7: Number of microsatellites where motif mismatches are either higher or lower**

445   **than the expected values of mismatches in different bird genomes.**

446

447   **Table S8: Genomic abundance of imperfect microsatellite loci based on length and number**

448   **of mismatches.**

449

450   **Table S9: Result for the relationship between average mismatches and body mass fitted in**

451   **PGLS analyses when different parameters were used to search microsatellites in the**

452   **genomes (minimum score = 15 and mismatch penalty = 3; minimum score = 10 and**

453   **mismatch penalty = 5). The average mismatches of imperfect microsatellites for the 65**

454   **birds are given on the next page below the result table.**

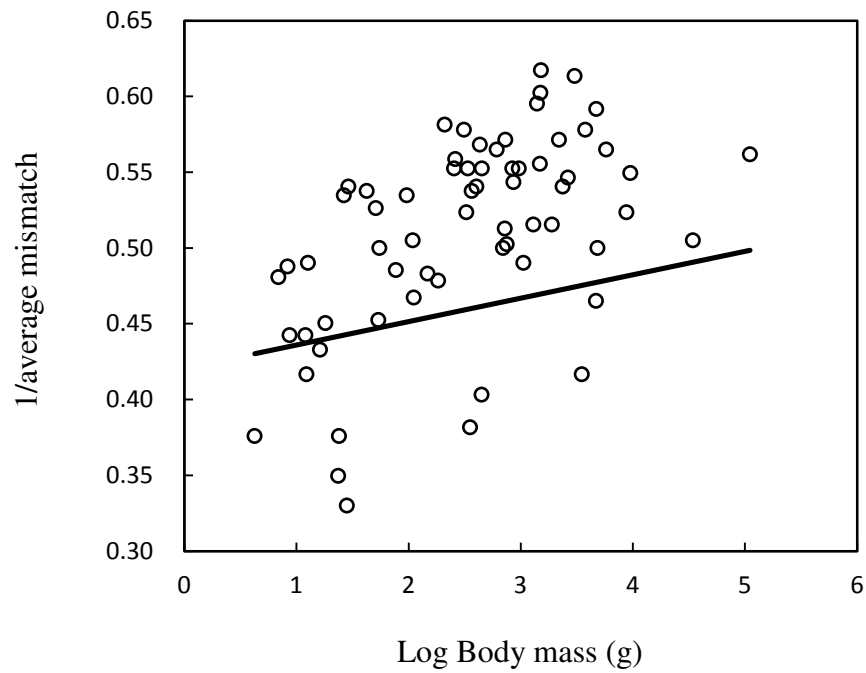
455

Result for the relationship between average mismatches and body mass fitted in PGLS analyses.

1   **Table 1**

Type	Model	R^2	Coefficients					
			Body mass			**GC content**		
			$\beta \pm SE$	t	**P**	$\beta \pm SE$	t	**P**
All	BM	0.282	0.015 ± 0.003	4.974	<0.001			
	BM+GC	0.340	0.015 ± 0.003	5.102	<0.001	-2.533 ±1.090	2.324	0.023
Di	BM	0.279	0.022 ± 0.005	4.932	<0.001			
	BM+GC	0.332	0.022 ± 0.004	5.033	<0.001	-2.940 ±1.255	2.342	0.022
Tri	BM	0.257	0.015 ± 0.003	4.664	<0.001			
	BM+GC	0.293	0.015 ± 0.003	4.711	<0.001	-2.044 ± 1.149	1.778	0.080
Tetra	BM	0.290	0.019 ±0.004	5.072	< 0.001			
	BM+GC	0.393	0.019 ±0.003	5.382	< 0.001	-4.074 ±1.257	3.241	0.002
Penta	BM	0.123	0.011 ±0.004	2.972	0.004			
	BM+GC	0.205	0.011 ±0.004	3.050	0.003	-3.272 ±1.296	2.525	0.014
Hexa	BM	0.254	0.013 ±0.003	4.634	< 0.001			
	BM+GC	0.268	0.013 ±0.003	4.620	< 0.001	-1.129 ±1.049	1.076	0.286

4   Key to symbols: All, all imperfect microsatellites; Di, Tri, Tetra, Penta, Hexa means imperfect
5   microsatellites with different repeat type; BM, Body mass; GC, GC content.

Regression scatterplot of the inverse of the average mismatches of imperfect SSRs on the log of body mass in whole genome scale.

Regression scatterplot of the inverse of the average mismatches of imperfect SSRs in five classes of repeat type on the log of body mass.